Title
Reply to ‘DNA methylation haplotypes as cancer markers’

Journal
Nature Genetics, 50(8)

ISSN
1061-4036

Authors
Diep, Dinh
Zhang, Kun

Publication Date
2018-08-01

DOI
10.1038/s41588-018-0186-9

Peer reviewed

Reply to 'DNA methylation haplotypes as cancer markers'

Diep and Zhang reply: We thank Greally and colleagues[1] for pointing out a number of inaccuracies in our publication[2] and the issue of potential overfitting, due to incomplete separation of test and training datasets. We have conducted a thorough internal audit and discovered that several components of the analysis indeed mixed training and test datasets and were therefore overfit to the published data. We acknowledge this as a major weakness of our published study. However, we would like to stress that our analysis based on methylation haplotype load and methylation haplotype blocks is reproducible and capable of deconvoluting heterogeneous tissue samples such as plasma DNA. To address the concern of overfitting, we performed an independent analysis of the data while carefully maintaining separation of training and test datasets. The results of this new analysis still support the three major conclusions of the publication[2]:

1. The haplotype-based metric methylation haplotype load (or MHL) is better able to identify tissue-specific differential methylation than the commonly used average methylation fraction (AMF) metric;
2. The tumor load of a plasma sample can be estimated using cancer tissue methylation markers;
3. The tissue of origin of a cancer plasma sample can be determined using tissue-specific methylation markers.

To ensure that our approach is accurate and transparent, we have released the code for the low-level sequence/haplotype processing and high-level statistical analysis and included the full data matrices for public release. The updated code and data matrices are being thoroughly tested and are available on GitHub (https://github.com/dinhdiep/MONOD2). Figure 1 illustrates the separation of training and test data for our analysis, and additional information on study design can be found in the Reporting Summary.

To show the utility of the MHL metric to identify differential methylation in an unbiased manner, we demonstrated the advantage of MHL with a published set of whole-genome bisulfite sequencing (WGBS) data (Schultz et al.[3]) and computed the AMF and MHL metrics in a previously identified set of differentially methylated regions (Lokk et al.[4]). As such, these regions were selected independently of the definitions of methylation haplotype block (MHB) and MHL. Using the group-specific index (GSI; a marker selection metric defined in ref. [2]) of the AMF values, we identified the top 150 differentially methylated regions associated with each tissue in the previously published marker set. We then plotted a heat map of the AMF and MHL values in these regions (Supplementary Fig. 1).

From the heat map, both the AMF and MHL metrics can be used to discriminate tissue types from one another; however, the ratio of the diagonal 'signal' methylation to the off-diagonal 'noise' methylation value is much higher in the MHL metric (Supplementary Fig. 2). The superior signal-to-noise ratio of the MHL metric allows for differential methylation to be more easily identified and used in samples with low signal (such as plasma).

We wanted to demonstrate that the MHL metric could be used to estimate the tumor load of plasma samples. We first performed pruning and *k*-nearest-neighbors (KNN) imputation on the MHL matrix, which removed samples with low coverage and imputed missing values. We were left with 30 colon cancer plasma samples, 29 lung cancer plasma samples and 69 normal plasma samples. For this analysis, we split the 69 healthy plasma samples into 'training' and 'test' sets; 46 samples were set aside for feature selection and training, while the remainder (23 samples) were used as a completely independent dataset to test the quantification (Supplementary Table 1). We next identified MHBs that were hypermethylated in lung and colon cancer tissue samples with respect to the training set using a one-sided *t* test (correcting for multiple testing with a false discovery rate (FDR) of 0.001 and applying a minimum difference of 0.3). These regions appeared to be hypomethylated in the test set of normal plasma samples and showed an intermediate methylation level in the plasma samples from patients with cancer (Supplementary Fig. 3). In fact, across these regions, the average MHL value for each group was significantly different between the plasma samples from cancer patients and the test set of normal plasma samples (two-sample *t* test, one-sided; colon cancer: $t = 7.4318$, degrees of freedom (d.f.) = 1,018, $P = 1.133 \times 10^{-13}$; lung cancer: $t = 8.8288$, d.f. = 1,834, $P = 2.2 \times 10^{-16}$) (Supplementary Fig. 4).

Next, we tested the ability of these cancer-associated blocks to quantify the tumor load of any given plasma sample. To calibrate the relationship between tumor load and the MHL values in these regions, we performed 20 sets of simulations in which we mixed sequencing reads from cancer tissue samples and normal plasma samples at a ratio of 1:5, 1:10, 1:20, 1:100 and 0:1 (for a total of 100 simulated datasets). We then computed the average MHL value in these regions for each simulation; as expected, the MHL value was highly correlated with tumor fraction (Supplementary Fig. 5). Finally, using the standard curve determined by simulations, we estimated the tumor load of each cancer plasma sample. As expected, the tumor load in the cancer plasma samples was statistically significantly higher than in the test normal plasma set (two-sample *t* test with unequal variance, one-sided; colon cancer: $t = 2.7338$, d.f. = 32.033, $P = 0.005055$; lung cancer: $t = 3.7686$, d.f. = 41.718, $P = 0.002547$) (Supplementary Fig. 6). This finding shows the ability to identify biological signals from plasma DNA using MHL values.

We further demonstrated that we could utilize tissue-specific differentially methylated regions identified from human normal tissue data to determine the tissue of origin of cancer plasma. To do this, we used an independent set of WGBS data from human tissues to identify a set of tissue-specific markers and then used these markers to classify each plasma sample to its tissue of origin. To be clear, the previous method also used WGBS tissue data to identify tissue-specific markers; however, additional features selection was performed with the plasma samples outside of the cross-validation loop of the original model-training process. Thus, while the previous set of markers demonstrated that tissue-of-origin mapping of plasma DNA was possible, these markers were difficult to generalize.

To overcome this weakness, we first identified a set of MHB regions that could be used to determine the tissue of origin of a sample. We used a set of training WGBS tissue data from the following nine tissues: GI (colon, small intestine, stomach,

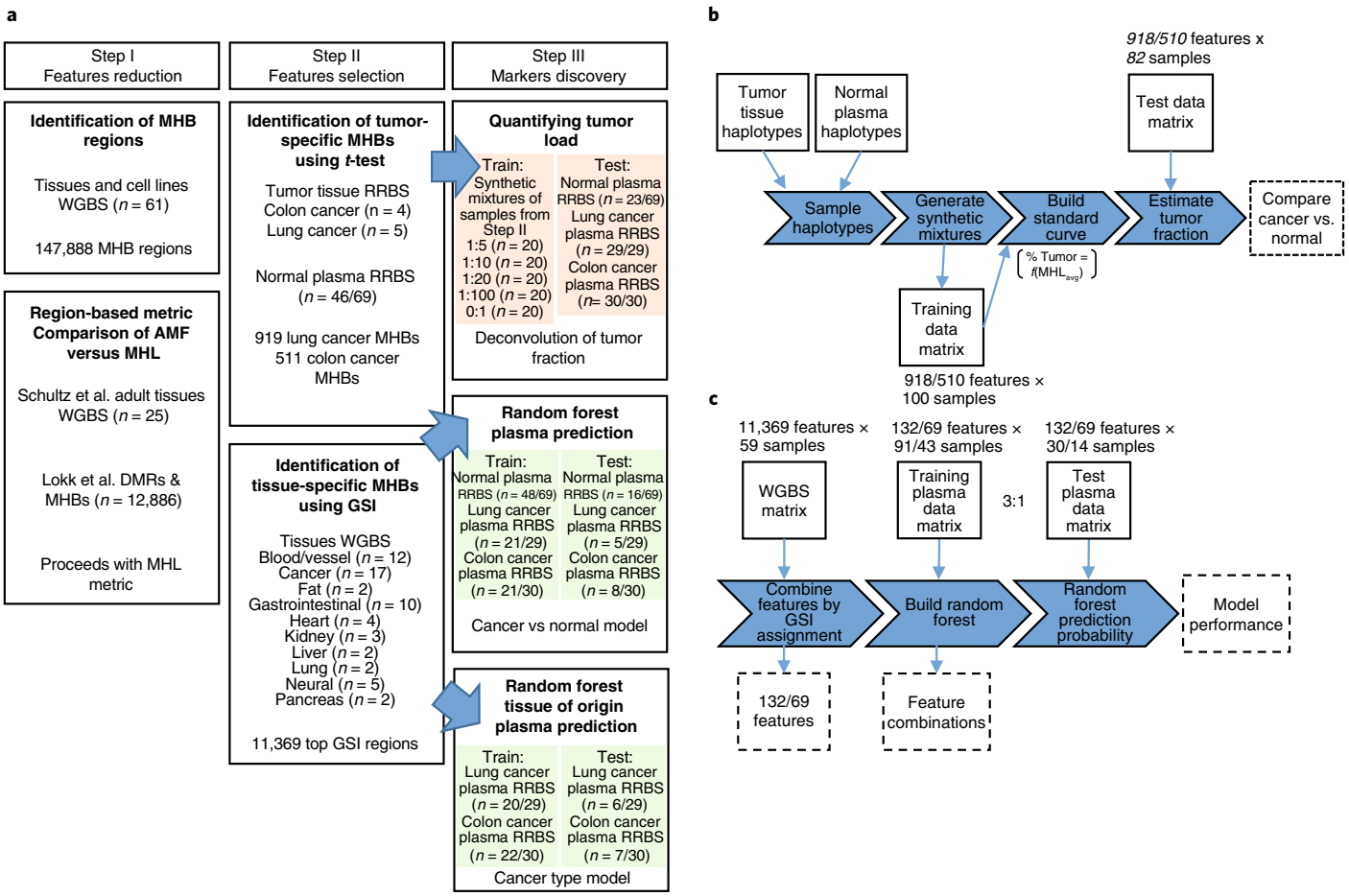

Fig. 1 | Marker discovery and validation workflow. a, The workflow is divided into three steps. The first step to define features (MHBs) was performed in ref. [2] with comparison of the AMF and MHL metrics using data from refs [3,4]. The second step to perform features filtering was carried out for identification of tumor-specific MHBs, tissue-specific MHBs, and colon- or lung-specific MHBs separately. The third step is to identify markers. **b**,**c**, The third and final step of marker discovery is performed either with synthetic mixing to generate a function that converts average MHL to tumor load (**b**) or binary classification using random forest (**c**).

esophagus), lung, neural, heart, liver, lung, pancreas, kidney and fat. For each MHB, we first calculated the haplotype load for the methylated haplotypes (MHL) and the unmethylated haplotypes (uMHL), with which we then calculated the GSI within this training set; MHBs with high GSI scores tended to be methylated or unmethylated in specific tissue types. We then selected a subset of 11,396 MHBs (Supplementary Table 2) that had (i) low signals in white blood cells, (ii) detectable signals in the nine normal tissue types or tumors, and (iii) a certain level of tissue specificity (GSI > 0.3 for cancer detection; GSI > 0.45 for tissue-of-origin classification). We want to emphasize that up to this point no plasma sample was used and hence all MHBs were solely selected and further weighted on the basis of the tissue specificity among human tissues, completely independent of plasma samples from cancer patients or healthy controls. We then generated a data matrix for all plasma samples on these MHBs. To reduce the number of features and avoid overfitting on small sample size, we took the weighted mean of all MHBs with similar tissue specificity and reduced a data matrix with 11,396 MHBs into a much smaller matrix with 132 tissue-specific features. For classification of tissues of origin, we further removed all cancer-related features and reduced the search space to 69 features related to the normal tissues.

To avoid overfitting in model building, we randomly sampled 75% of plasma samples to create a training dataset from the reduced-representation bisulfite sequencing (RRBS) plasma samples (90 for cancer versus normal classification; 42 for colon versus lung classification). The remaining 25% of samples (29 for cancer versus normal classification; 13 for colon versus lung classification) were held out as a test dataset (Supplementary Table 1). We performed the testing in two stages: first, with a binary classifier (i.e., testing the ability of the model to identify a plasma sample as either 'normal' or 'cancer') and second, with a tissue-of-origin classifier separating colon cancer plasma from lung cancer plasma samples.

By using a random forest–based binary classifier, we were able to distinguish normal from cancer samples at an area under the ROC curve (AUC) of 0.83 and lung cancer from colon cancer at an AUC of 0.76 (Supplementary Fig. 7) in the 25% of samples held out for testing. To further evaluate stability, we repeated the random splitting of training and test samples 50 times and determined the interquantile range (IQR) for cancer versus normal classification as [0.71, 0.84] and for lung cancer versus colon cancer classification as [0.64, 0.83]. Note that our models

were derived on very limited sample sizes during the training stage. These prediction accuracies represent a proof of concept for the methylation-haplotype-based approach and should be further improved when more samples are available for model building.

In addition to the analysis presented above, we also want to take the opportunity to address a few points raised by Greally and colleagues[1]. We appreciate their comments on the supplementary deconvolution analyses we presented in Supplementary Table 7a–d of ref. [2]. It is correct that only the samples with a whole blood (WB) fraction above 0.3 were selected in the summary table, as specified in Supplementary Table 7a. The cutoff of 0.3 allowed separation of a distribution that looked bimodal when we generated a plot similar to that in Fig. 1 from Greally and colleagues[1]. As shown in Supplementary Fig. 8a of ref. [2], the deconvolution accuracy was only high when the fraction to estimate was low. While all the information was clearly presented in the table, we apologize for not making this explicit in the main text. Note that the motivation for this particular analysis was to create a benchmark with a previously published method[5] and to document aspects where the method seemed to have worked. This analysis is completely independent of the second deconvolution analysis presented in Supplementary Figs. 3–6 and in Fig. 4 of ref. [2] and of the detection of tumor and tissue of origin in plasma presented in Supplementary Fig. 7 and in Fig. 5 of ref. [2].

We respectfully disagree with the comment on the biological significance of MHBs. First, we would like to stress that there was no technical error in the calculation of enrichment or the MHL metric. Second, the motivation for developing the concept of methylation haplotypes and MHBs was to investigate regions in the human genome that exhibit differences in CpG methylation patterns from one cell type (or cell state) to another. The genomic regions that show static methylation patterns (either fully methylated or unmethylated across all cell or tissue types) are not covered by MHBs on the basis of our definition; these regions are not related to cell-type-specific regulation and are therefore not of interest in this study (and most other studies concerning epigenetic variation and regulation). Third, MHBs and the MHL metric allow us to capture CpG methylation trends that are coordinated locally along single DNA molecules. This is related to de novo methylation or demethylation mechanisms and the processivity of the related enzymes; one hypothesis on how local methylation patterns can be established is that a methyltransferase or demethylase enzyme can reach a CpG site, add or remove a methyl group, and then slide along the DNA or 'hop' locally to process nearby CpG sites. The enrichment of MHBs in variably methylated regions (VMRs) favors this hypothesis.

Code availability. Full data matrices and codes can be accessed at GitHub (https://github.com/dinhdiep/MONOD2).

Data availability. The raw sequencing data can be accessed through the Gene Expression Omnibus (GEO) with accession GSE79279. ❐

Dinh Diep and Kun Zhang*
Department of Bioengineering, University of California at San Diego, La Jolla, CA, USA.
**e-mail: kzhang@bioeng.ucsd.edu*

Published online: 27 July 2018
https://doi.org/10.1038/s41588-018-0186-9

References

1. Seoighe, C., Tosh, N. J. & Greally, J. M. *Nat. Genet.* https://doi.org/10.1038/s41588-018-0185-x (2018).
2. Guo, S. et al. *Nat. Genet.* **49**, 635–642 (2017).
3. Schultz, M. D. et al. *Nature* **523**, 212–216 (2015).
4. Lokk, K. et al. *Genome Biol.* **15**, r54 (2014).
5. Sun, K. et al. *Proc. Natl. Acad. Sci. USA* **112**, E5503–E5512 (2015).

Author contributions

D.D. and K.Z. performed all analyses and wrote the manuscript.

Competing interests

D.D. and K.Z. were listed as co-inventors on patent applications related to this study. K.Z. is a co-founder and scientific advisor of Singlera Genomics.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0186-9.

nature research

Corresponding Author: Kun Zhang

Date: 5/15/2018

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work we publish. This form is published with all life science papers and is intended to promote consistency and transparency in reporting. All life sciences submissions use this form; while some list items might not apply to an individual manuscript, all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

We determined that with a significance level of 0.05, sample sizes of 23 and 30, and power of 0.8, we can detect large effect sizes (~0.7-0.8) for two groups using the T-Test.

2. Data exclusions

Describe any data exclusions.

Some plasma data were excluded in tumor load estimation because they had low coverages/many missing values. Then other samples were determined to have abnormally high genomic DNA levels which could interfere with the signal from cell free DNA, and thus, they were excluded.

3. Replication

Describe whether the experimental findings were reliably reproduced.

Experimental findings were reliably reproduced.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

We used randomization to determine whether the results would be able to generalize. Therefore, we used 70-80% of the data selected using randomization softwares for training and the remainder for testing.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Investigators were not blinded to group allocation during analysis.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or the Methods section if additional space is needed).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☒	☐	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly.
☒	☐	A statement indicating how many times each experiment was replicated
☐	☒	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒	The test results (e.g. *p* values) given as exact values whenever possible and with confidence intervals noted
☐	☒	A summary of the descriptive statistics, including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	The code were deposited in GitHub at: https://github.com/dinhdiep/MONOD2 (The data from this publication was generated using "v2" branch).

For all studies, we encourage code deposition in a community repository (e.g. GitHub). Authors must make computer code available to editors and reviewers upon request. The *Nature Methods* guidance for providing algorithms and software for publication may be useful for any submission.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	No unique material was used.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	No antibody was used.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	No cell line was used.
b. Describe the method of cell line authentication used.	No cell line was used.
c. Report whether the cell lines were tested for mycoplasma contamination.	No cell line was used.
d. If any of the cell lines used in the paper are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	No cell line was used.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	No animal was used.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	Not applicable.